Running Head: FALSE-POSITIVE PSYCHOLOGY

False-Positive Psychology: *Undisclosed* Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant[a]

Joseph P. Simmons
Yale University

Leif D. Nelson
University of California, Berkeley

Uri Simonsohn
University of Pennsylvania

In press at *Psychological Science*

Word Count: 4,040

[a] Authors contributed equally. Author order is alphabetical (controlling for father's age [reverse coded]).

Abstract

This paper accomplishes two things. First, we show that despite our field's nominal endorsement of a low rate of false-positive findings ($p \leq .05$), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. It involves six concrete requirements for authors and four guidelines for reviewers, imposing a minimal burden on the publication process.

Our job as scientists is to discover truths about the world. We generate hypotheses, collect data, and examine whether or not they are consistent with those hypotheses. Although we aspire to always be accurate, errors are inevitable.

Perhaps the most costly error is a *false-positive*, the incorrect rejection of a null hypothesis. First, false-positives are particularly persistent. Because null results have many possible causes, failures to replicate previous findings are never conclusive. Furthermore, because it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive to even attempt them. Second, false-positives waste resources: They inspire investment in fruitless research programs and can lead to ineffective policy changes. Finally, a field known for publishing false-positives risks losing its credibility.

We show that despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p \leq .05$), current standards for disclosing details of data collection and analyses make false-positives vastly more likely. In fact, it is unacceptably easy to publish “statistically-significant” evidence consistent with *any* hypothesis.

The culprit is a construct we refer to as *researcher degrees-of-freedom*. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined and/or transformed?

It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted) for researchers to explore various analytic alternatives, to search for a combination that yields “statistical significance,” and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

This exploratory behavior is not the by-product of *malicious intent*, but rather the result of two factors: (1) *ambiguity* in how best to make these decisions and (2) the researcher's *desire* to find a statistically significant result. A large literature documents that people are self-serving in their interpretation of ambiguous information and remarkably adept at reaching justifiable conclusions that mesh with their desires (Babcock & Loewenstein, 1997; Dawson, Gilovich, & Regan, 2002; Gilovich, 1983; Hastorf & Cantril, 1954; Kunda, 1990; Zuckerman, 1979). This literature suggests that when we as researchers face ambiguous analytic decisions, we will tend to conclude, with convincing (self) justification, that the appropriate ones are those that result in $p \leq .05$.

Importantly, ambiguity is rampant in empirical research. As an example, consider a very simple decision faced by researchers analyzing reaction-times: how to treat outliers. In a perusal of roughly 30 *Psychological Science* articles, we discovered considerable inconsistency in, and hence considerable ambiguity about, this decision. Most (but not all) researchers excluded some responses for being too fast, but what constituted "too fast" varied enormously: the fastest 2.5%, or faster than 2 *SD* from the mean, or faster than 100 or 150 or 200 or 300 ms. Similarly, what constituted "too slow" varied enormously: the slowest 2.5% or 10%, or 2 or 2.5 or 3 *SD* slower than the mean, or 1.5 *SD* slower from that condition's mean, or slower than 1000 or 1200 or 1500 or 2000 or 3000 or 5000 ms. None of these decisions is necessarily incorrect, but that fact makes *any* of them justifiable and hence potential fodder for self-serving justifications.

How Bad Can It Be? A Demonstration of *Chronological Rejuvenation*

To help illustrate the problem, we conducted two experiments designed to demonstrate something false: that certain songs can change listeners' age. Everything reported below actually happened.[1]

Study 1: Musical Contrast and Subjective Age

In Study 1, we investigated whether listening to a children's song induces an age contrast, making people feel older.

In exchange for payment, 30 University of Pennsylvania undergraduates sat at computer terminals, donned headphones, and were randomly assigned to listen to either a control song ("Kalimba," an instrumental song that comes free with the Windows 7 operating system) or a children's song ("Hot Potato," performed by The Wiggles).

After listening to part of the song, participants completed an ostensibly unrelated task: They answered "How old do you feel right now?" by choosing among *very young, young, neither young nor old, old,* and *very old*. They also reported their father's age, allowing us to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: People felt older after listening to "Hot-Potato" than after listening to the control song (adjusted Ms = 2.54 vs. 2.06), $F(1, 27) = 5.06$, $p = .033$.

In Study 2, we sought a conceptual replication and extension. Having demonstrated that listening to a children's song makes people feel older, Study 2 investigated whether listening to a song about older age makes people *actually* younger.

Study 2: Musical Contrast and Chronological Rejuvenation

Twenty undergraduates drawn from the same pool as Study 1 listened to either "When I am 64" by The Beatles or "Kalimba." In an ostensibly unrelated task, they then indicated their birth

[1] Our goal was to pursue a research question that would not implicate any particular brand of research. Our concerns apply to all branches of experimental psychology, and to the other sciences as well.

date (mm/dd/yyyy) and their father’s age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: People were nearly a year-and-a-half younger after listening to “When I am 64” rather than “Kalimba” (adjusted *Ms* = 20.1 vs. 21.5), *F*(1, 17) = 4.92, *p* = .033.

Discussion

These two studies were conducted with real participants, employed legitimate statistical analysis, and are reported truthfully. Nevertheless, they seem to support hypotheses that are unlikely (Study 1) or necessarily false (Study 2).

Before detailing the researcher degrees-of-freedom we employed to achieve these “findings,” we provide a more systematic analysis of how researcher degrees-of-freedom influence statistical significance. Impatient readers can consult Figure 3.

How Bad Can It Be? Simulations

In this section, we describe computer simulations of experimental data that estimate how researcher degrees-of-freedom influence the probability of a false-positive result. They assess the impact of four common degrees-of-freedom, involving flexibility in (1) choosing among dependent variables, (2) sample size, (3) use of covariates, (4) reporting subsets of experimental conditions, and (5) their combination.

We generated random samples with each observation independently drawn from a normal distribution, performed sets of analyses on each sample, and observed how often at least one of the resulting p-values in each sample was below standard significance levels.

For example, imagine a researcher who collects two dependent variables, say liking and willingness-to-pay. S/he can test whether the manipulation affected liking, whether the manipulation affected willingness-to-pay, and whether the manipulation affected a combination of liking and willingness-to-pay. The likelihood that one of these tests produces a significant result is at least *somewhat* more likely than .05. We conducted 15,000 simulations of this scenario (and others) to estimate the size of "somewhat."[2]

We report the results of our simulations in Table 1. Row *a* shows that flexibility in analyzing two dependent variables (correlated at $r = .50$) nearly doubles the probability of obtaining a false-positive finding.[3]

In row *b*, we consider a researcher who collects 20 observations per condition and then tests for significance. If the result is significant, she stops collecting data and reports the result. If the result is non-significant, she collects 10 additional observations per condition, and then again tests for significance. Row *b* shows that this seemingly small degree-of-freedom increases the false-positive rate by about 50%.

Row *c* shows the effect of flexibility in controlling for gender or for an interaction between gender and the independent variable.[4] Such flexibility leads to a false-positive rate of 11.7%. Row *d* shows that running three conditions (e.g., low, medium, high) and reporting the results for any two or all three (e.g., low vs. medium; low vs. high; medium vs. high; low vs. medium vs. high) generates a false-positive rate of 12.6%.

Rows *e-g* consider combinations of those reported above, with bottom row *g* reporting the false-positive rate if the researcher uses all of these degrees-of-freedom, leading to a stunning

[2] We conducted simulations instead of deriving close-form solutions because the combinations of researcher degrees-of-freedom we considered lead to fairly complex derivations without adding much insight over simulation results.

[3] The lower the correlation between the two dependent variables, the higher the false-positive rate produced by considering both. Intuitively, if $r = 1$, then both are the same variable; if $r = 0$, then the two tests are entirely independent.

[4] We assigned each observation a gender of 1 or 0, independently, with 50% probability; "gender" is a placeholder for any covariate with similar properties.

61%! A researcher is more likely than not to falsely detect a significant effect if s/he uses just these four common researcher degrees-of-freedom.

As high as these estimates are, they may actually be conservative. We did not consider many other degrees-of-freedom researchers commonly use, including testing and choosing among more than two dependent variables (and the various ways to combine them), testing and choosing among more than one covariate (and the various ways to combine them), excluding subsets of participants and/or trials, flexibility in deciding whether early data was part of a pilot study or part of the experiment proper, etc.

A Closer Look at Flexibility in Sample Size

Researchers often decide when to stop data collection based on interim data analysis. Notably, a recent survey of behavioral scientists found that about 70% of them admitted to having done so (John, Loewenstein, & Prelec, 2011). In conversations with colleagues, we have learned that many believe this practice exerts no more than a trivial influence on false-positive rates.

Contradicting this intuition, Figure 1 reports the false-positive rates from additional simulations for a researcher who has already collected either 10 or 20 observations within each of two conditions, and then tests for significance every 1, 5, 10, or 20 per-condition observations after that. The researcher stops collecting data either once statistical significance is obtained or when n = 50.[5]

The figure shows that a researcher who starts with n = 10 and then tests for significance after every new per-condition observation finds a significant effect 22% of the time. Figure 2 depicts an illustrative example continuing sampling until n = 70. It plots p-values from t-tests conducted after each pair of observations. It contradicts the often-espoused yet erroneous intuition that if an

[5] We always use lowercase n to denote *per-condition* sample size.

effect is significant with a small sample size then it would necessarily be significant with a larger one.

Solution

In this section, we offer our solution to the flexibility/ambiguity problem, in the form of six requirements for authors and four guidelines for reviewers (see Table 2). This solution substantially mitigates the problem while imposing only a minimal burden on authors, reviewers, and readers. Our solution leaves the right and responsibility of identifying the most appropriate way to conduct research in the hands of researchers, requiring only that authors provide appropriately transparent descriptions of their methods so that reviewers and readers can make informed decisions regarding the credibility of their findings. We assume that the vast majority of researchers strive for honesty; this solution will not help in the unusual case of willful deception.

Requirements for Authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article. This may mean reporting the outcome of power calculations, but also of arbitrary rules such as “we decided to collect 100 observations” or to “collect as many observations as we could before the end of the semester.” The rule itself is secondary, but it must be determined ex-ante and be reported.

2. Authors must collect at least 20 observations per cell, or else provide a compelling cost-of-data-collection justification. This offers extra protection for the first requirement. Samples smaller than 20 per cell are simply not powerful enough to detect most effects, and so there is usually no good reason to decide in advance to collect such a small number of observations. Smaller samples, it follows, are much more likely to reflect interim data analysis and a flexible termination rule. In addition, as Figure 2 shows, larger minimum sample sizes can lessen the impact of violating requirement 1.

3. Authors must list all variables collected in a study. This prevents researchers from reporting only a convenient subset of the many measures that were collected, allowing readers and reviewers to easily identify possible researcher degrees-of-freedom. Because authors are required to just *list* those variables, this requirement increases paper length by a few words per otherwise shrouded variable. We encourage authors to begin the list with “only,” to assure readers the list is exhaustive (e.g., “participants reported *only* their age and gender”).

4. Authors must report all experimental conditions, including failed manipulations. This prevents authors from selectively choosing only to report the condition comparisons that are consistent with their hypothesis. As with the previous requirement, we encourage authors to include the word “only” (e.g., “participants were randomly assigned to one of *only* three conditions”).

5. If observations are eliminated, authors must also report the statistical results if they are included. This makes transparent the extent to which a finding is reliant on the exclusion of observations, puts appropriate pressure on authors to justify the elimination of data, and encourages reviewers to explicitly consider whether such exclusions are warranted. Correctly interpreting a finding may require some data exclusions; this requirement is merely designed to draw attention to those results that hinge on ex-post decisions about which data to exclude.

6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate. Reporting covariate-free results makes transparent the extent to which a finding is reliant on their presence, putting appropriate pressure on authors to justify it, and encourages reviewers to consider whether it is warranted. Some findings may be persuasive even if covariates are required for their detection, but we should place greater scrutiny on results that do hinge on covariates despite random assignment.

Guidelines for Reviewers

7. Reviewers should ensure that authors follow the requirements. Review teams are the gatekeepers of the scientific community, and should encourage authors to not only rule out alternative explanations, but to more convincingly demonstrate that their findings are not due to chance alone. This means prioritizing transparency over tidiness; if a wonderful study is partially marred by a peculiar exclusion or an inconsistent condition, those imperfections should be retained. If reviewers require authors to follow these requirements, they will.

8. Reviewers should be more tolerant of imperfections in results. One reason researchers exploit researcher degrees-of-freedom is the unreasonable expectation we often impose as reviewers for every data pattern to be (significantly) as predicted. Underpowered studies with perfect results are the ones that should invite extra scrutiny.

9. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions. Even if authors follow all guidelines listed above, they will necessarily still face arbitrary decisions. For example, should they subtract the baseline level of the dependent variable from the final result or should they use it as a covariate? When there is no obviously correct way to answer questions like this, the reviewer should ask for alternatives. For example, reviewer reports might include questions like, "Do the results also hold if the baseline measure is instead used as a covariate?" Similarly, reviewers should ensure that arbitrary decisions are used consistently across studies (e.g., "Do the results holds for Study 3 if gender is entered as a covariate, as done in Study 2?").[6] If a result holds only for one arbitrary

[6] It is important that these alternatives be reported in the manuscript (or in an appendix) rather than merely in a private response to reviewers, so that the research community has access to the results.

specification, then those involved have learned a great deal about the robustness (or lack thereof) of the effect.

10. If data collection or analysis justifications are not compelling, reviewers should require the authors to conduct an exact replication. If a reviewer is not persuaded by the justifications for a given researcher degree-of-freedom and/or results from a robustness check, the reviewer should ask the author to conduct an *exact* replication of the study and its analysis. We realize that this is a costly solution and it should be used selectively; however, "never" is too selective.

The Solutions In Action: Revisiting Chronological Rejuvenation

To show how our solutions would work in practice, we return to our Study 2, which "showed" that people get younger when listening to The Beatles, and we report it again in Figure 3, following the requirements above. The merits of reporting transparency should be evident, but a few highlights are worth mentioning.

First, notice that our original reporting redacted the (many) measures other than father's age that we collected (including the dependent variable from Study 1, feelings of oldness). A reviewer would hence have been unable to assess the flexibility involved in selecting it as a control. Second, by reporting results without the covariate, readers are more likely to discover its critical role in achieving a significant result. Combined with the full list of variables now disclosed, reviewers would have an easy time asking for robustness checks, such as "Do the results from Study 1 replicate in Study 2?" (They do not: $F(1, 17) = 2.07$, $p = .168$, opposite direction). Finally, notice that we did not determine the study's termination rule in advance, instead monitoring statistical significance approximately every 10 observations. Moreover, our sample size did not reach the $n = 20$ threshold set by our requirements.

The redacted version of the study we reported above fully adheres to currently acceptable reporting standards and is, not coincidentally, deceptively persuasive. The version reported in Figure 3 would be—appropriately—all but impossible to publish.

General Discussion

Criticisms

Criticism of our solution comes in two flavors: It does not go far enough and it go too far.

Not Far Enough. Our solution does not lead to the disclosure of all degrees-of-freedom. Most notably, it cannot reveal those arising from reporting only experiments that "work" (i.e., the file-drawer problem). One might address this by requiring researchers to submit all studies to a public repository, whether or not they are "successful" (see e.g. Ioannidis, 2005; Schooler, 2011). Although we are sympathetic to this suggestion, it does come with significant practical challenges: How is submission enforced? How does one ensure that study descriptions are understandably written and appropriately classified? Importantly, in order for the repository to be effective, it must adhere to our disclosure policy, for it is impossible to interpret study results, whether successful or not, unless researcher degrees-of-freedom are disclosed. The repository is an ambitious long-term extension of our recommended solution, not a substitute.

A reviewer of our paper worried that our solution may not go far enough because authors have "tremendous disincentives" to disclose exploited researcher degrees-of-freedom. Although researchers obviously have incentives to publish, if editors and reviewers enforce our solution, authors will have even stronger incentives to accurately disclose. Our solution turns inconsequential sins of omission (leaving out inconvenient facts) into consequential, potentially career-ending sins of commission (writing demonstrably false statements). Journals implementing our disclosure requirements will create a virtuous cycle of transparency and accountability that eliminates the disincentive problem.

Too Far. Alternatively, someone may be concerned that our guidelines prevent researchers from conducting exploratory research. What if researchers do not know which dependent measures will be sensitive to the manipulation, for example, or how such dependent measures should be scored or combined? We all should of course engage in exploratory research, but we should be required to either report it as such (i.e., following the six requirements), or to complement it with (and possibly only report) confirmatory research consisting of *exact* replications of the design and analysis that "worked" in the exploratory phase.

Non-Solutions

In the process of devising our solution, we considered a number of alternative ways to address the researcher degrees-of-freedom problem. All other solutions, we believe, are less practical, less effective, or both. This does not necessarily mean they are not worth pursuing, as they may be useful for other reasons, but they do not in our view address the problem of researcher degrees-of-freedom.

Correcting alpha levels. As done with multiple-hypothesis testing, one may consider adjusting the critical level, α, *a la* Bonferroni, as a function of the number of researcher degrees-of-freedom employed in each study. Something like this has been proposed for medical trials that monitor outcomes as the study progresses (see e.g. Pocock, 1977).

First, given the broad and ambiguous set of degrees-of-freedom in question, it is unclear which and how many of them contribute to any given finding, and hence what their effect is on the false-positive rate. Second, unless there is an explicit rule about exactly how to adjust α for each degree-of-freedom, *and their combination* (see bottom rows in Table 2), the additional ambiguity may make things worse by introducing new degrees-of-freedom.

We have a similar reaction to calls for using Bayesian rather than frequentist approaches to analyzing experimental data (see, e.g., Wagenmakers, Wetzels, Borsboom, & Van der Maas, 2011). Although the Bayesian approach has many virtues, it *increases* researcher degrees-of-freedom. First, it offers a new set of analyses (in addition to all frequentists ones) that authors could flexibly try out on their data. Second, Bayesian statistics require making additional judgments (e.g., the prior distribution) on a case-by-case basis, providing yet more researcher degrees-of-freedom.

Conceptual replications. Because conceptual replications, in contrast to exact replications, do not bind researchers to make *the same* analytic decisions across studies, they are unfortunately misleading as a solution to the problem at hand. In a paper with a conceptual replication, for instance, authors may choose two of three conditions in Study 1 and report one measure, while choosing a different pair of conditions and a different measure in Study 2. Indeed, that is what we did in the experiments reported above.

Posting materials and data. We are strongly supportive of all journals requiring authors to make their original materials and data publicly available. However, this is not likely to address the problem of interest, as it imposes too high a cost on readers and reviewers to examine, in real time, the credibility of a particular claim. Readers should not need to download data, load it into their statistical packages, and start running analyses to learn the importance of controlling for father's age, or need to read pages of additional materials to learn that the researchers had simply dropped the "Hot Potato" condition.

Furthermore, if a journal allows the redaction of a condition from the report, for example, it would presumably also allow its redaction from the raw data and “original” materials, making the entire transparency effort futile.

Concluding Remarks

Our goal as scientists is not to publish as many articles as we can, but to discover and disseminate truth. Many of us – and this includes the three of us – often lose sight of this, yielding to the pressure to do whatever is justifiable to compile a set of studies that we can publish. This is not driven by a willingness to deceive but from the self-serving interpretation of ambiguity, which enables us to convince ourselves that whichever decisions produced the most publishable outcome must have also been the most appropriate. This article advocates a set of disclosure requirements that imposes minimal costs on authors, readers, and reviewers. These solutions will not rid us of publication pressures, but they will limit what we are able to justify as acceptable to others *and* to ourselves. We should embrace these disclosure requirements as if the credibility of our profession depended on them. Because it does.

References

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives, 11*, 109-126.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin, 28*, 1379-1387.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44*, 1110-1126.

Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *Journal of Abnormal and Social Psychology, 49*, 129-134.

Ioannidis, J. P. A. (2005). Why most published research findings are false. [Editorial Material]. *Plos Medicine, 2*, 696-701.

John, L., Loewenstein, G. F., & Prelec, D. (2011). *Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling*. Working paper.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*, 191-199.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*, 437-437.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426-432.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality, 47*, 245-287.

Author Note

Joseph P. Simmons, Yale School of Management; Leif D. Nelson, Haas School of Business, University of California, Berkeley; Uri Simonsohn, The Wharton School, University of Pennsylvania.

We thank Jon Baron, Jason Dana, Victor Ferreira, Geoff Goodwin, Jack Hershey, Dave Huber, Hal Pashler, and Jonathan Schooler for their valuable comments. Address correspondence to uws@wharton.upenn.edu, leif_nelson@haas.berkeley.edu, or joseph.simmons@yale.edu.

Table 1

Likelihood of Obtaining a False-Positive Result

Row	Degree(s)-of-freedom	Significance level		
		p<.1	p<.05	p<.01
a	Two dependent variables (correlated r=.5)	17.8%	9.5%	2.2%
b	Adding exactly 10 more observations per cell	14.5%	7.7%	1.6%
c	Controlling for gender or gender interaction with treatment	21.6%	11.7%	2.7%
d	Dropping (or not) 1 of 3 conditions	23.2%	12.6%	2.8%
e	Combine a,b	26.0%	14.4%	3.3%
f	Combine a,b,c	50.9%	30.9%	8.4%
g	Combine a,b,c,d	81.5%	60.7%	21.5%

Note. Numbers correspond to the percentage of (15,000) simulated samples in which at least one of a set of analyses is significant. Observations are drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Row *a* considers three t-tests, one on each of two dependent variables and a third on their average. Row *b*: a t-test performed with 20 observations per cell and another with an additional 10. Row *c*: a difference of means, an ANCOVA with a gender main effect and one with a gender interaction (each observation is assigned to female = 1 with 50% probability). Row *d*: three possible pairings of three conditions and a linear trend for all three jointly.

Table 2

Simple Solutions

Requirements For Authors

1. Authors must decide their rule for terminating data collection before data collection begins and report this rule in the article
2. Authors must collect at least 20 observations per cell, or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report the statistical results if they are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Guidelines for Reviewers

7. Reviewers should ensure that authors follow the requirements.
8. Reviewers should be more tolerant of imperfections in results.
9. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.
10. If data collection or analysis justifications are not compelling, reviewers should require the authors to conduct an exact replication.

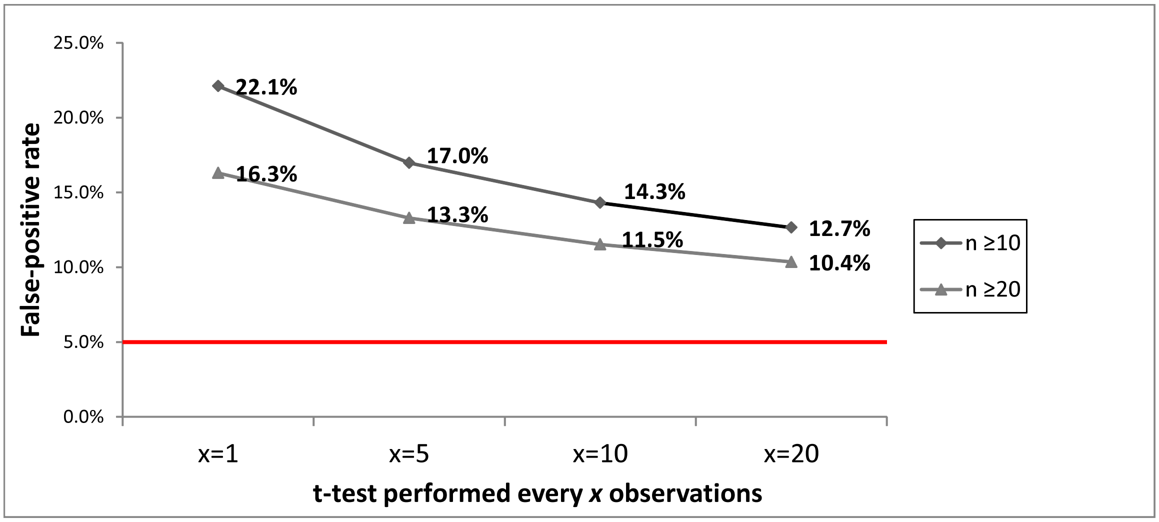
Figure Captions

Figure 1. Likelihood of obtaining a falsely significant result when data collection ends upon obtaining significance ($p \leq .05$).

Figure 2. Illustrative simulation of t-tests performed on samples of increasing size.

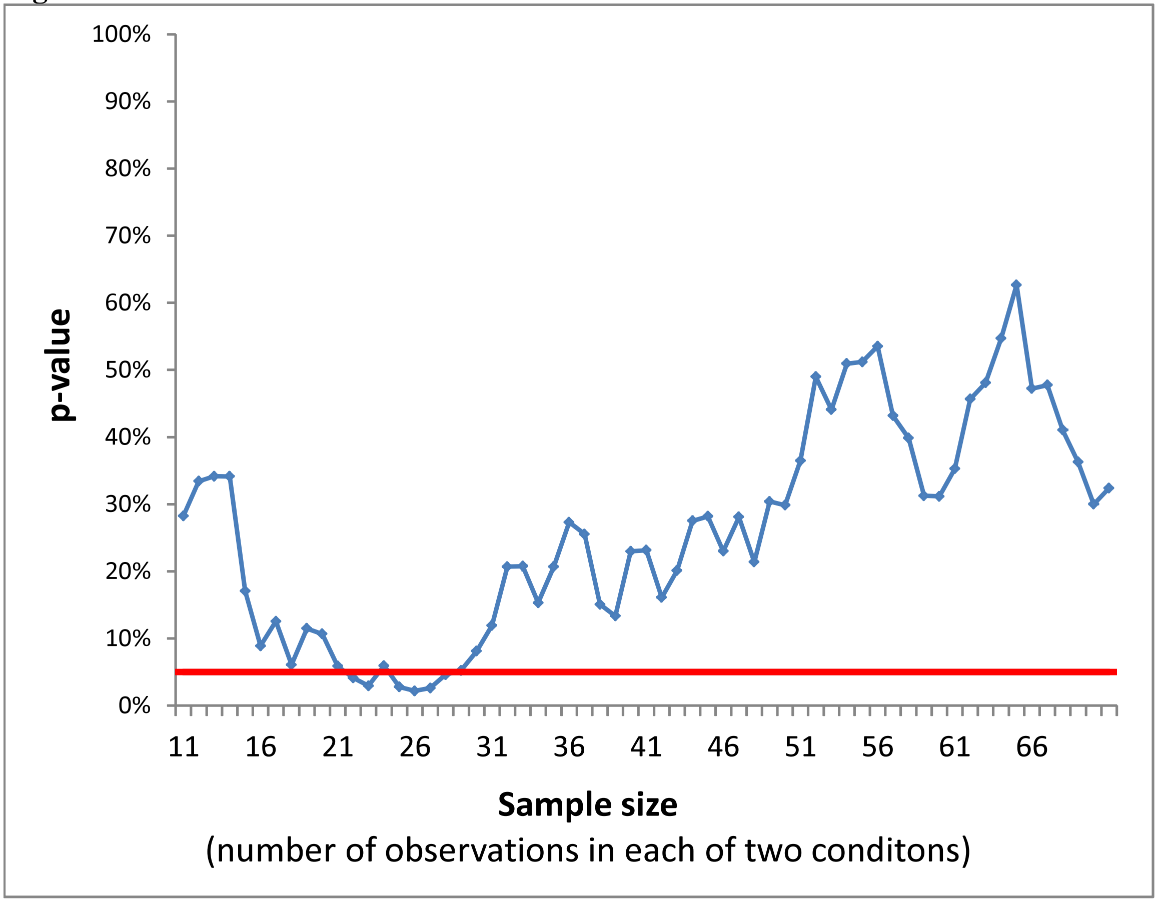
Figure 3. The difference between what we reported and what our solutions would require us to report.

Figure 1



Note. Calculations are based on 15,000 simulations of two conditions with observations drawn from a normal distribution. A t-test comparing means is conducted every x = 1, 5, 10, 20 observations per-cell, starting at *n* =10 or *n* = 20 per cell. The false-positive rate is the percentage of simulations for which at least one of the t-tests reveals $p \leq .05$ (two-tailed).

Figure 2



Note. We simulated a two-condition experiment with *n* = 70 by drawing each observation independently from a normal distribution. The graph plots the two-tailed p-values obtained when comparing the two samples after each pair of observations, starting at *n* = 11 and ending at *n* = 70. This particular simulation was chosen for illustrative purposes; it is not meant to be representative. No single draw of random data is.

Figure 3

Study 2. Original report (vs. Requirement-Compliant Report)

~~Twenty~~ Thirty-four **undergraduates drawn from the same pool as Study 1 listened** only **to either "When I am 64" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles**.** We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **In an ostensibly unrelated task, they then indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, agreement with "computers are complicated machines," **their father's age,** their mother's age, whether they would take advantage of an early bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days", and their gender, **to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: People were nearly a year-and-a-half younger after listening to "When I am 64" rather than "Kalimba" (adjusted *M*s = 20.1 vs. 21.5), *F*(1, 17) = 4.92, *p* = .033. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively, *F*(1, 18) = 1.01, *p* = .33).